APPLICATION

FOR

UNITED STATES LETTERS PATENT

APPLICANT NAME: Batra et al.

TITLE: SYSTEM AND METHOD FOR SECURING GENOMIC INFORMATION

DOCKET NO.: CHA920040003US1

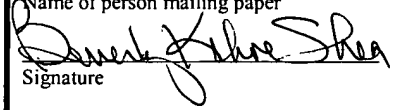
INTERNATIONAL BUSINESS MACHINES CORPORATION

CERTIFICATE OF MAILING UNDER 37 CFR 1.10

I hereby certify that, on the date shown below, this correspondence is being deposited with the United States Postal Service in an envelope addressed to Mail Stop Patent Application, Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450 as "Express Mail Post Office to Addressee"
Mailing Label No. EV 393 299 693 US

on April 1, 2004

Beverly Kehoe Shea
Name of person mailing paper



Signature

April 1, 2004
Date

SYSTEM AND METHOD FOR SECURING GENOMIC INFORMATION

BACKGROUND OF THE INVENTION

1. Technical Field

The present invention relates to securing genomic information, and more specifically relates to a system and method for selectively securing genetic coding regions being communicated over a network using web services.

2. Related Art

Grid computing (or the use of a *computational grid*) is a term of art for applying the resources of many computers in a network to a single problem at the same time - usually to a scientific or technical problem that requires a great number of computer processing cycles or access to large amounts of data.

In one important application, grid computing technologies enable the sharing of bioinformatics data from different sites by creating a virtual organization of the data. Specifically, bioinformatics grids allow the sharing of geographically distributed bioinformatics data. Thus, genetic research results can be stored on a local system and shared with the research community immediately. Moreover, users no longer need to know the location of their target information, but are able to access and retrieve data in a transparent manner. This paradigm is extremely appropriate for many types of bioinformatics research efforts, including large-scale genomic and proteomic activities.

Grid technologies are feasible thanks in part to a standardized network technology referred to as web services. Web services (sometimes called *application services*) are network services that are made available from an application server for web users or other web-connected programs. The use of web services is a major web trend for communicating data and services on the Internet. Because web services can be implemented on a peer-to-peer basis, and not just on a central server, it lends itself to grid computing.

Standardized data exchange within web services is enabled with the use of Extensible Markup Language (XML) documents. In a typical bioinformatics application, XML documents are utilized to hold important information, such as nucleotide chains and the identification of genetic sequences, which are communicated remotely to the computational grid.

While the use of web services and computational grids provide numerous advantages when applied to bioinformatics, there are several challenges that remain. One of the challenges with using web services for bioinformatics relates to security. Existing secure web service standards only provide encryption mechanisms for either specific attributes of the XML message, or the entire XML message. However, because nucleotide chains are very large, e.g., it is not unusual for a chain to comprise many megabytes, encrypting and decrypting the entire chain requires a significant amount of computational time. For example, the magnaporthe grisea genome has approximately 40 millions basepairs, and the length of a human genome is approximately 3,000 million basepairs. Accordingly, a need exists for a system for handling and selectively securing regions of bioinformatics sequences being transmitted and processed in a web services environment.

SUMMARY OF THE INVENTION

The present invention addresses the above-mentioned problems, as well as others, by providing a system and method for providing security to a nucleotide chain over a network by encrypting only selected regions of the chain. In a first aspect, the invention provides a security system for securing an electronic transmission of a nucleotide chain, comprising: a system for identifying coding and non-coding regions in the nucleotide chain; and a system for selectively encrypting only the coding regions identified in the nucleotide chain.

In a second aspect, the invention provides a method for securely transmitting a nucleotide chain, comprising: identifying coding and non-coding regions in the nucleotide chain; selectively encrypting only the coding regions identified in the nucleotide chain to generate encrypted coding regions and unencrypted non-coding regions; and transmitting the encrypted coding regions and unencrypted non-coding regions.

In a third aspect, the invention provides a program product stored on a recordable medium for encoding a nucleotide chain, comprising: means for identifying coding and non-coding regions in the nucleotide chain; and means for selectively encrypting only the coding regions identified in the nucleotide chain.

In a fourth aspect, the invention provides a program product stored on a recordable medium for decoding an encoded nucleotide chain, comprising: means for identifying coding and non-coding regions in the encoded nucleotide chain; means for selectively decrypting only the coding regions identified in the encoded nucleotide chain; and means for reassembling the coding and non-coding regions to generate a decoded nucleotide chain.

BRIEF DESCRIPTION OF THE DRAWINGS

These and other features of this invention will be more readily understood from the following detailed description of the various aspects of the invention taken in conjunction with the accompanying drawings in which:

Figure 1 depicts a bioinformatics system for securely communicating an encoded nucleotide chain in accordance with the present invention.

Figure 2 depicts an encrypting system in accordance with the present invention.

Figure 3 depicts a decrypting system in accordance with the present invention.

DETAILED DESCRIPTION OF THE INVENTION

Referring now to the drawings, Figure 1 depicts a bioinformatics system 11 for communicating an encoded nucleotide chain 12 from a first application 10 to a second application 20. In one exemplary embodiment, applications 10 and 20 provide a query or search system wherein application 10 provides a remote application for inputting a nucleotide chain query, and application 20 provides a bioinformatics database, which can be queried with the inputted nucleotide chain. In other possible embodiments, applications 10 and 20 may represent any two systems that communicate bioinformatics data, for example, applications 10 and 20 may represent nodes within a computational grid, a system for uploading bioinformatics information to a database, a client and server, two servers, email applications, etc.

In the embodiment depicted in Figure 1, data transfer between applications 10 and 20 is implemented using web services 16, which utilizes one or more XML documents 24 to transmit the encoded nucleotide chain 13. Application 10 includes a nucleotide chain encrypting system 14 for encrypting portions of chain 12, and application 20 includes a decrypting system 18 for

decrypting an encoded chain 13. Obviously, either or both applications 10 and 20 may include both an encrypting and a decrypting system to facilitate secure bi-directional data transfers.

As noted above, encrypting an entire nucleotide chain can be a computationally intensive process. To address this, the present invention encrypts only selected coding regions of the nucleotide chain 12. Coding regions, i.e., exons, are the only part of a nucleotide chain that convey information about the genome being studied. The non-coding regions or introns, represent junk DNA that do not convey information about the genome. In accordance with the present invention, the non-coding regions are not encrypted, thereby greatly reducing the computational requirements of bioinformatics system 11. As shown in Figure 1, an encoded chain 13 is generated using XML document(s) 24 comprising encrypted chain data 26 (comprised of coding regions) and unencrypted chain data 28 (comprised of non-coding chain data).

Referring now to Figures 2 and 3, exemplary embodiments of nucleotide chain encrypting system 14 and nucleotide chain decrypting system 16 are described in further detail. As shown in Figure 2, nucleotide chain encrypting system 14 includes a coding region identification system 29 that receives a nucleotide chain 12 and identifies the coding regions 30 and non-coding regions 32 in the chain 12. Systems for handling this process are well known in the art, and are therefore not explained in further detail herein. Coding region identification system 30 splits nucleotide chain 12 into "islands" of coding and non-coding regions 30, 32, e.g.,

[non-coding region][coding region][non-coding region][coding region] ....

The coding regions 30 are encrypted, in this case, using cipher block chain (CBC) encryption system 34. CBC is known encryption technique that encrypts a sequence of bits as a single unit, or block, with a cipher key. CBC uses a chaining mechanism that allows the decryption of a

block of ciphertext to depend on all the preceding ciphertext blocks. Thus, the validity of a block is contained in the immediately previous cipertext block. Accordingly, the validity of each coding region can be proved by the immediately preceding coding region. While CBC is a particularly robust solution for this type of application, it should be recognized that any encryption, encoding, or security technique could be utilized to secure the coding regions 30, and thus fall within the scope of this invention.

XML document packaging system 36 receives the encrypted coding regions 30 and unencrypted non-coding regions 32, and "packages" the regions in one or more XML documents 24. The regions can be packaged in any manner, e.g., each region could be stored into a unique XML document; multiple regions could be stored in a single XML document; multiple regions could be stored in multiple XML documents, etc. It should be understood that nucleotide chain encrypting system 14 describes one exemplary embodiment for encrypting and packaging coding and non-coding regions 30, 32, and that other embodiments are possible and fall within the scope of the invention. For instance, nucleotide chain encrypting system 14 could package the regions 30, 32 into one or more XML documents before the coding regions 30 are encrypted.

The following is an exemplary XML document containing coding and non-coding regions of a simplified nucleotide sequence:

CGATCCAA...CAG**AGTCCA**GGACCCAA...ATG**AAACGTCCATT**

wherein the bolded nucleotides indicate coding regions, and "..." indicates nucleotides omitted for brevity purposes.

****

<XML doc>doc 1</XML doc>

<Sequence Name>Nuc Seq 1</Sequence Name>

<non-coding_region.1>CGATCCAA...CAG </non-coding_region.1>

<coding_region.1>AGTCCA</coding_region.1>

<non-coding_region.2>GGACCCAA...ATG </non-coding_region.2>

<coding_region.2>AAACGTCCATT</coding_region.2>

****

In the above example, coding_region.1 and coding_region.2 are encrypted to secure the exact coding sequences that convey information about the genome being studied. As noted, using CBC, the validity of coding_region.2 can be proved based on coding_region.1. Non-coding_region.1 and non-coding_region.1 are not encrypted since they do not convey any relevant information. Obviously, the exact format (e.g., tag names, etc.) of XML document(s) 24 can be implemented in any workable/desirable manner.

Referring to Figure 3, nucleotide chain decrypting system 16 is shown, which is used to regenerate nucleotide chain 12 from XML document(s) 24. Nucleotide decrypting system 16 includes an XML document parsing system 46, which identifies the encrypted coding regions 42 and unencrypted coding regions 44. Encrypted coding regions 42 are subsequently decrypted by CBC decryption system 40. Once decrypted, chain reassembly system 48 reassembles the regions back to the original nucleotide chain 12.

It is understood that the systems, functions, mechanisms, methods, engines and modules described herein can be implemented in hardware, software, or a combination of hardware and software. They may be implemented by any type of computer system or other apparatus adapted for carrying out the methods described herein. A typical combination of hardware and software could be a general-purpose computer system with a computer program that, when loaded and executed, controls the computer system such that it carries out the methods described herein.

Alternatively, a specific use computer, containing specialized hardware for carrying out one or more of the functional tasks of the invention could be utilized. The present invention can also be embedded in a computer program product, which comprises all the features enabling the implementation of the methods and functions described herein, and which - when loaded in a computer system - is able to carry out these methods and functions. Computer program, software program, program, program product, or software, in the present context mean any expression, in any language, code or notation, of a set of instructions intended to cause a system having an information processing capability to perform a particular function either directly or after either or both of the following: (a) conversion to another language, code or notation; and/or (b) reproduction in a different material form.

The foregoing description of the invention has been presented for purposes of illustration and description. It is not intended to be exhaustive or to limit the invention to the precise form disclosed, and obviously, many modifications and variations are possible. For instance, while the present invention has been described with reference to a system utilizing XML documents, the concepts and techniques could be applied to any system for communicating electronic data. Such modifications and variations that may be apparent to a person skilled in the art are intended to be included within the scope of this invention as defined by the accompanying claims.